Les Actes du BRG, 6 (2006) 41-55

Article original

Un modèle géostatistique pour la détection et la localisation des discontinuités génétiques spatiales entre populations

Jean-François COSSON[(1)*], Arnaud ESTOUP[(1)], Aurélie COULON[(2)], Maxime GALAN[(1)], Frédéric MORTIER[(3)], A.J. Marc HEWISON[(2)], Gilles GUILLOT[(4)]

[(1)] INRA, Centre de Biologie et de Gestion des Populations, Campus International de Baillarguet, CS 30 016, F34988 Montferrier-sur-Lez, France
[(2)] INRA, Comportement et Écologie de la Faune Sauvage, BP 52627, 31326 Castanet-Tolosan Cedex, France
[(3)] CIRAD, Département Forêt, 34398 Montpellier, France
[(4)] INRA Mathématique et Informatique Appliquées, INA P-G, 16, rue Claude Bernard, 75231 Paris Cedex 5, France

Abstract: A spatial statistical model for landscape genetics. Landscape genetics is a new discipline that aims to provide information on how landscape and environmental features influence population genetic structure. This approach is of primary importance in population management and conservation biology because it provides valuable knowledge about landscape connectivity. The first key step of landscape genetics is the spatial detection and location of genetic discontinuities between populations. However, efficient methods for achieving this task are lacking. In this research project, we first clarify what is conceptually involved in the spatial modelling of genetic data. Then we describe a Bayesian model that allows inference of the location of such genetic discontinuities from individual georeferenced multilocus genotypes, without a priori knowledge on the number of populational units and their limits. In this method, the global set of sampled individuals is modelled as a spatial mixture of panmictic populations, and the spatial organization of populations is modelled through coloured Voronoi tessellation. In addition to spatially locating genetic discontinuities, the method quantifies the amount of spatial dependence in the data set, estimates the number of populations in the studied area, assigns individuals to their population of origin, and detects migrants between populations. The performance of the method was evaluated through the analysis of simulated data sets. Results showed good performances for standard microsatellite data sets (*e.g.*, 100 individuals genotyped at 10 loci with 10 alleles per locus), with high but also low levels of population differentiation ($F_{ST} < 0.05$). The method was then applied to two real data sets on large mammals with contrasted differentiation levels. The first application, to wolverines (*Gulo gulo*) sampled in the North-western United States, showed the ability of the method to detect populations consistent with

* Correspondance et tirés à part : cosson@ensam.inra.fr

landscape structures known to slow down dispersal movements of that species, and to locate putative migrants in a context of rather high genetic differentiation (F_{ST} from 0.08 to 0.17). The second application, to roe deer (*Capreolus capreolus*) in South-western France, illustrate the ability of the method to infer genetic discontinuities coherent with landscape structures (highways, canals) in a situation of very low genetic differentiation (F_{ST} = 0.008). A computer program named GENELAND is freely available at http://www.inapg.inra.fr/ens_rech/mathinfo/personnel/guillot/Geneland.html. A mailing list for users is managed by one of us (G. Guillot).

barriers/ Bayesian computations/ gene flow/ landscape connectivity/ landscape genetics

Résumé : La génétique du paysage est une nouvelle discipline dont le but est de décrire l'influence des structures paysagères et environnementales sur la structuration spatiale de la variabilité génétique. Cette approche est de première importance pour la gestion des populations et en biologie de la conservation car elle donne des informations pertinentes sur la connectivité des habitats. La première étape de la génétique du paysage est la détection et la localisation dans l'espace des discontinuités génétiques entre populations. Pourtant, il n'existait pas jusqu'à présent de méthode d'analyse efficace pour atteindre ce but. Dans ce programme, nous avons clarifié les concepts liés à la modélisation spatiale des données génétiques. Nous avons ensuite décrit un modèle Bayésien permettant d'inférer les discontinuités génétiques à partir de génotypes multilocus individuels géoréférencés, sans à priori sur le nombre de populations et leurs limites spatiales. Dans cette méthode, le jeu de données global (génétique et spatial) est modélisé comme un mélange de populations panmictiques, dont l'organisation spatiale est modélisée par des cellules de Voronoi. Outre la localisation des discontinuités génétiques, la méthode quantifie le degré de dépendance spatiale dans le jeu de données, estime le nombre de populations dans l'aire d'étude, assigne les individus à leur population d'origine, et détecte les éventuels migrants entre ces populations. La performance de cette méthode a été évaluée grâce à l'analyse de données simulées. Les résultats ont montré de bonnes performances pour des jeux de données standards à des locus microsatellites (une centaine d'individus génotypés à 10 locus avec 10 allèles chacun), y compris pour des niveaux de différenciation relativement faibles (F_{ST} < 0,05). Cette méthode a ensuite été appliquée à deux jeux de données réels sur des grands mammifères avec des niveaux de différenciation très différents. La première application, sur le glouton (*Gulo gulo*) dans l'Ouest des États-Unis, montre la capacité de cette méthode à détecter des populations cohérentes avec les structures paysagères connues pour freiner la dispersion chez cette espèce, et à localiser un certain nombre de migrants potentiels dans un contexte de différenciation génétique assez élevée (F_{ST} de 0,08 à 0,17). La deuxième application, sur le chevreuil dans le Sud-Ouest de la France, illustre la capacité de cette approche à inférer des discontinuités génétiques cohérentes d'un point de vue paysager (autoroutes, canaux), dans un contexte de très faible différenciation génétique (F_{ST} = 0,008). Un programme informatique, dénommé GENELAND, est disponible gratuitement à : http://www.inapg.inra.fr/ens_rech/mathinfo/personnel/guillot/Geneland.html. Une mailing liste a également été mise en place et compte une soixantaine d'utilisateurs enregistrés.

barrières/ statistiques Bayésiennes/ flux de gènes/ connectivité du paysage/ génétique du paysage

1. INTRODUCTION

La génétique du paysage est une nouvelle discipline dont le but est de décrire l'influence des structures paysagères et environnementales sur la structuration spatiale de la variabilité génétique [1]. La localisation de discontinuités génétiques et la corrélation de ces discontinuités avec les éléments du paysage procurent des informations essentielles pour de nombreuses disciplines scientifiques, de la biologie évolutive à la biologie de la conservation. Cette approche est de première importance pour la gestion des populations car elle donne des informations pertinentes sur la connectivité des paysages qui peut être définie comme le degré auquel un paysage facilite ou gêne les déplacements d'organismes entre les tâches de ressources [2]. En particulier, elle permet de définir de façon objective des unités de gestion et/ou de conservation [3] qui sont jusqu'à présent, et faute de méthodologie adaptée, généralement définies de façon intuitive et/ou correspondent, pour des raisons pratiques, à des contraintes administratives. L'étude des flux de gènes et de leurs discontinuités est également indispensable pour prédire ou simuler dans des situations réelles l'évolution de la diversité génétique et de la dynamique démographique des espèces sous différents scénarios, notamment la fragmentation et/ou la modification des paysages sous la pression anthropique. L'inventaire de la diversité génétique nécessaire à l'élaboration de toute stratégie de conservation/gestion des ressources génétiques doit donc prendre en compte autant que faire se peut la possibilité de détecter les différentes entités génétiques présentes dans un espace géographique donné.

Une étape clef de la génétique du paysage est la détection et la localisation de discontinuités génétiques entre les populations [1]. Idéalement cette étape devrait être basée sur des méthodes qui ne posent aucune hypothèse à priori sur les limites entre populations. Cela implique que l'individu soit l'unité opérationnelle pour l'inférence. Plusieurs méthodes récentes permettent de grouper des individus dans des unités populationnelles et de détecter des migrants entre ces unités, sans hypothèse à priori sur les limites populationnelles [4], [5], [6], [7], [8]. Ces méthodes ne prennent cependant pas en compte de façon explicite la dimension spatiale. Les coordonnées géographiques des échantillons ne sont utilisées qu'une fois l'inférence sur les populations réalisée, afin de visualiser leurs contours sur une carte [9], [10]. Cette approche n'est pas vraiment satisfaisante si on considère que les populations sont généralement organisées dans l'espace, ce qui constitue une information qui n'est pas exploitée si on ne prend pas en compte cette dimension spatiale. Cette observation est à l'origine du travail présenté. Nous

proposons d'expliciter un modèle statistique qui incorpore directement les localisations spatiales des individus dans l'inférence des unités populationnelles, avec l'optique que ce postulat pertinent d'un point de vue biologique augmente la sensibilité de l'inférence.

Nos principaux objectifs dans cet article sont i) de présenter une méthode géostatistique originale spécialement adaptée à la nature discrète des données de génotypes multilocus, qui permette de définir sans à priori le nombre et les limites géographiques des unités populationnelles présentes dans un paysage. Cette approche repose sur l'analyse de génotypes individuels géoréférencés dans un espace à deux dimensions ; ii) de tester les performances de cette méthode sur des jeux de données simulés notamment en fonction de la différenciation génétique entre les unités populationnelles ; enfin iii) d'appliquer cette méthode sur des grands mammifères dans des paysages où la présence de barrières aux flux de gènes est suspectée pour des raisons écologiques.

2. FORMALISATION DU MODÈLE STATISTIQUE

D'un point de vue statistique, la question que nous traitons consiste à regrouper les individus dans un certain nombre de populations sur la base du génotype multilocus et de la localisation spatiale des individus. On suppose que l'on a observé n individus diploïdes, on connaît leurs génotypes ainsi que leurs coordonnées spatiales à certaines précisions près (on peut considérer le cas d'animaux non sédentaires, à condition d'avoir une information sur leurs déplacements potentiels). On suppose que ces n individus proviennent de K populations différentes, K étant inconnu. On souhaite modéliser la structure génétique spatiale de cet échantillon. En statistique, ce terme désigne un écart au désordre total modélisé par des variables aléatoires indépendantes. Dans notre contexte, nous proposons de distinguer la structuration à trois niveaux différents : au niveau de l'organisation spatiale des populations, au niveau des fréquences alléliques dans les différentes populations et au niveau des génotypes à l'intérieur de chaque population. L'idée centrale qui guide notre travail est qu'en général, les populations animales sont spatialement structurées, au sens où deux individus géographiquement proches ont une probabilité plus grande d'appartenir à la même population que deux individus géographiquement éloignés. On souhaite pouvoir injecter cette information dans le modèle de manière qualitative, et quantifier l'intensité de cette dépendance spatiale. Pour cela nous avons recours à un modèle développé initialement en géostatistique pour modéliser les formations géologiques des gisements miniers ou des réservoirs pétroliers [11].

Nous supposons que les aires de répartition de chaque population peuvent se décomposer en un nombre fini de polygones de Voronoi engendrés

par un processus de Poisson sur le domaine d'étude (fig. 1). Les fréquences alléliques de chaque population sont calculées selon deux modèles disponibles au choix : suivant des lois de Dirichlet indépendantes d'une population à l'autre et d'un locus à l'autre (modèle D, inspiré de Pritchard *et al.* [4]), ou suivant des lois de Dirichlet corrélées entre elles (modèle F, inspiré de Falush *et al.* [6]. Nous supposons qu'à l'intérieur de chaque population, les génotypes individuels sont à l'équilibre de Hardy-Weinberg et qu'il n'y a pas de déséquilibre de liaison. Autrement dit, conditionnellement à la partition de l'espace et aux fréquences alléliques, il y a indépendance de toutes les variables observées. Les individus d'une même population sont « conditionnellement indépendants », c'est-à-dire qu'une fois les paramètres estimés, il ne reste plus de dépendance résiduelle : les frontières entre populations expliquent « totalement » l'écart à l'indépendance observé. Cette hypothèse peut être assez restrictive dans un contexte de populations naturelles (nombreuses sources possibles d'écart à l'équilibre de Hardy-Weinberg), mais constitue une simplification nécessaire dans ce modèle déjà assez complexe.

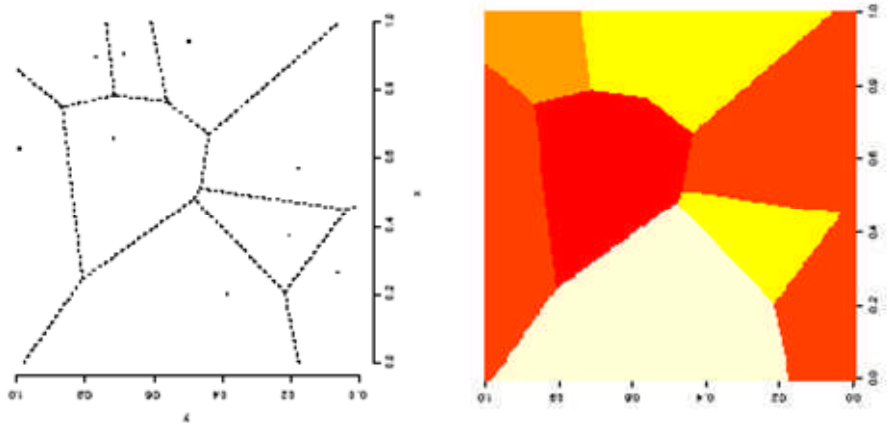

Figure 1 : Illustration du modèle statistique proposé pour le partitionnement de l'espace géographique. Chaque population est supposée occuper un territoire. Les frontières de ce territoire sont des inconnues du problème. Nous supposons que chaque territoire peut être décomposé en une réunion de polygones convexes engendrés par un semis de points distribués au hasard dans le domaine (selon une loi de Poisson). Cette hypothèse a surtout pour but de permettre un paramétrage simple des territoires de chaque population et n'admet pas forcément une interprétation écologique. La figure montre une simulation où on suppose que l'espace est occupé par quatre populations distinctes. La figure à gauche montre les points poissonniens et les polygones de Voronoi qu'ils engendrent, la figure à droite les mêmes polygones après affectation à une population. La probabilité d'appartenance de deux individus à une même population dépend alors de leur localisation, et donc de la distance géographique qui les sépare (à la différence de tous les modèles existants, où cette probabilité ne dépend que des génotypes des indi-

vidus). La position et la couleur (*i.e* la population d'appartenance) de ces polygones doivent être estimées à partir des données.

Enfin nous prenons en compte le fait que les coordonnées spatiales des individus fournies au moment de l'échantillonnage ne sont pas forcément les coordonnées les plus pertinentes à prendre en compte (déplacement au moment de la capture, animaux non sédentaires, erreurs de positionnement...). C'est pourquoi nous introduisons une différence entre des coordonnées vraies et des coordonnées observées reliées par une équation du type *coord. observées* = *coord. vraies* + *erreur*, la distribution statistique de l'erreur dépendant fortement du contexte.

Les paramètres qui interviennent dans ce modèle sont : K ; m ; u ; c ; f_A ; d ; f ; s où K est le nombre de populations, m est le nombre de polygones, (u_1,..., u_m) sont les coordonnées des centres des polygones, (c_1 ,…, c_m) sont les variables de classes (codant l'appartenance de chaque polygone à une population), f_{Alj} sont les fréquences alléliques dans la population ancestrale (au locus l, pour l'allèle j), (d_1,…, d_K) sont les dérives, f_{klj} sont les fréquences alléliques dans la population k (au locus l, pour l'allèle j) et (s_1,…, s_n) sont les vraies coordonnées spatiales des individus, considérées comme inconnues. Les paramètres sont estimés en spécifiant des distributions à priori sur chaque bloc de paramètres et en simulant une chaîne de Markov dont la loi asymptotique est la distribution à posteriori du vecteur de paramètres. On réalise la simulation par un algorithme à sauts réversibles adapté de l'article de Green [11]. Pour une description complète du modèle statistique intitulé GENELAND, voir Guillot *et al.* [12], [13].

3. APPLICATION SUR DONNÉES SIMULÉES

Une caractéristique importante de notre modèle est sa capacité à estimer un nombre inconnu de populations (K). La précision de cette estimation a été testée avec de nombreux jeux de données simulées pour différentes valeurs de K (Guillot *et al.* [12] pour les détails). Les résultats montrent que GENELAND obtient d'excellents résultats pour des niveaux de différenciation variables (F_{ST} de 0,01 à 0,30) avec le modèle D. En revanche, l'utilisation du modèle F aboutit à une surestimation systématique de K, qui est d'autant plus importante que la différenciation génétique est faible.

La capacité du modèle à assigner les individus à leur population d'origine a été également testée avec de nombreux jeux de données simulées pour des niveaux de différenciation génétique et des degrés de dépendance spatiale variables (détails dans Guillot *et al.* [12]). Les résultats indiquent que les deux modèles de fréquences alléliques (modèles D et F) donnent des résultats similaires, et que notre modèle géostatistique est toujours plus performant

que les modèles non spatiaux équivalents. Ceci est d'autant plus vrai que la dépendance spatiale est forte (m petit) et le niveau de différenciation faible ($F_{ST} < 0{,}04$).

Enfin, nous avons testé notre modèle sur sa capacité à localiser correctement les frontières entre les populations. La figure 2 montre les sorties graphiques pour deux jeux de données simulées avec un fort ($F_{ST} = 0{,}16$) et un faible ($F_{ST} = 0{,}01$) niveau de différenciation génétique. Chaque jeu de données est constitué de 200 individus appartenant à deux populations (100 dans chaque population) typés à 10 locus avec 10 allèles par locus.

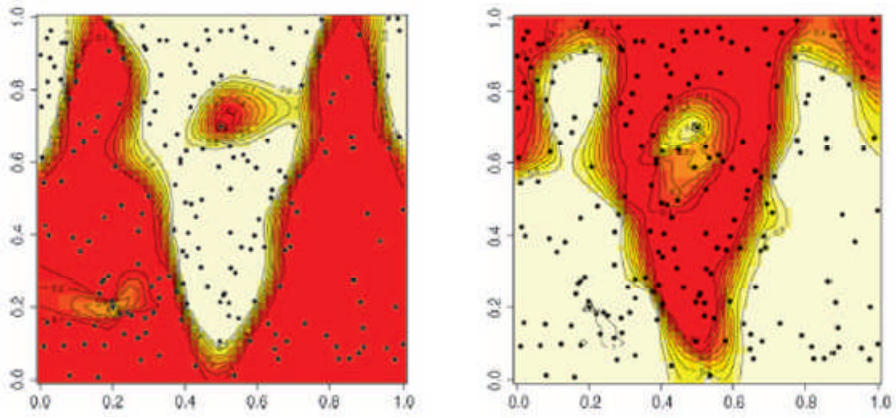

Figure 2 : Illustration de la capacité de notre modèle à classer des individus à partir de deux jeux de données simulées. Deux populations panmictiques séparées par une frontière sinusoïdale sont simulées (10 loci, 10 allèles par locus), pour deux niveaux de différenciation génétique (à droite $F_{ST} = 0{,}16$; à gauche $F_{ST} = 0{,}01$). Les individus sont positionnés aléatoirement au sein du domaine spatial de la population à laquelle ils appartiennent, et un individu de chaque population est placé au hasard de l'autre côté de la frontière par rapport à sa population d'origine. Le graphique montre la probabilité à posteriori de chaque pixel du domaine d'appartenir à une des deux populations et les aires des deux populations sont représentées en coloration claire ou foncée. Les flèches désignent les deux individus migrants également attendus. La frontière est parfaitement détectée ainsi que la présence des deux migrants, avec une précision toutefois inférieure pour le plus faible niveau de différenciation génétique. (Dessiné d'après [12]).

Les deux populations sont séparées par une barrière sinusoïdale, et les individus sont répartis aléatoirement dans le domaine spatial de la population à laquelle il appartient. Les analyses révèlent que la frontière entre les populations est très correctement localisée, avec une précision croissant avec le niveau de différenciation. D'autre part, des simulations ont également démontré la capacité de GENELAND à identifier correctement les migrants de première génération (fig. 2).

4. APPLICATION AUX GLOUTONS DU MONTANA

Nous avons analysé un jeu de données préalablement publié par Cegelsky *et al.* [9] sur le glouton (*Gulo gulo*), un carnivore de taille moyenne, largement répandu en Amérique du Nord. Les gloutons sont des animaux très mobiles, avec des distances de dispersion qui peuvent dépasser 300 km pour une année, mais ils sont également très sensibles aux modifications paysagères. Quatre-vingt neuf individus échantillonnés dans un paysage du Montana très fragmenté par les activités humaines ont été typés à 10 microsatellites. En utilisant des procédures de classification, notamment STRUCTURE [4], GENECLASS [14] et la procédure itérative de Vázquez-Dominguez *et al.* [15], Cegelsky et ses collaborateurs ont apporté des preuves d'une structuration géographique en trois groupes très différentiés (valeurs de F_{ST} de 0,08 à 0,10).

Nous avons réanalysé ces données avec GENELAND de la façon suivante : 10 simulations indépendantes de 200 000 itérations en permettant à K (le nombre de populations) de varier librement entre 1 et 15, intensité maximale du processus de Poisson fixée à 100, modèle D pour l'estimation des fréquences alléliques. Nous avons ensuite estimé le nombre de populations par la valeur modale de K dans ces 10 simulations indépendantes (K = 6), puis lancé cinq simulations MCMC (Monte-Carlo par Chaînes de Markov) avec K fixé à cette valeur sans toucher aux autres paramètres (plusieurs itérations ont été effectuées afin de vérifier la concordance des estimations). La probabilité d'appartenance à posteriori de chaque pixel du domaine spatial a été calculée pour chacune de ces cinq simulations (après une courte période de chauffe, *i.e. burning period*), de même que l'assignation des individus aux populations inférées.

Deux des six populations sont « vides », dans la mesure où elles ne sont les plus probables pour aucun des individus. Nous n'avons pas d'explication pour ces populations « fantômes » qui sont régulièrement rencontrées avec les jeux de données réelles, et semblent se localiser dans les aires géographiques peu (ou pas) échantillonnées (cf. discussion dans [12], [13] et [16]). Notre analyse avec GENELAND donne de fortes indications pour une population supplémentaire à celles déterminées précédemment par Cegelsky *et al.* [9], située au nord d'une de ces dernières. Les trois autres populations occupent des domaines spatiaux similaires à ceux déterminés par les approches non spatiales utilisées dans la publication initiale (fig. 3). Le principal gain de l'approche spatiale semble donc être une plus grande sensibilité. De façon remarquable, un examen minutieux des données cartographiques (*landcover*) de la zone d'étude (fig. 2 de [9]) révèle l'existence d'habitats anthropisés sur la nouvelle frontière proposée par notre analyse (entre populations #3 et #6, fig. 3). On note également la détection de migrants probables qui sont pour la plupart communs à ceux détectés dans la publication initiale.

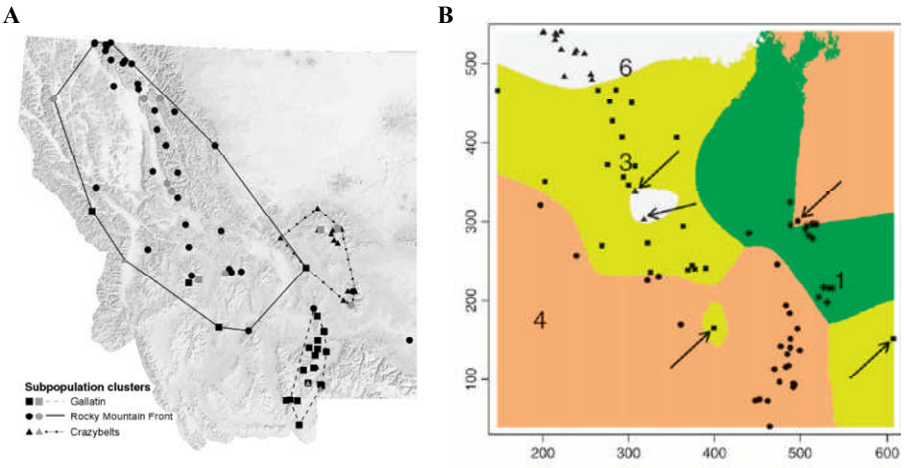

Figure 3 : Structuration génétique du glouton (*Gulo gulo*) dans le Montana, USA. (A) Unités génétiques proposées par STRUCTURE [4] et la méthode itérative de [15]. Les aires ont été dessinées autour de chaque groupe par la méthode des polygones convexes. Les symboles gris clair indiquent les individus pour lesquels les deux méthodes sont en désaccord. (B) Unités génétiques et barrières proposées par GENELAND. Les populations 2 et 5 non représentées sont des populations « fantômes » (voir le texte pour les détails). On remarque une unité génétique supplémentaire dans le Nord de la zone, qui est soutenue par la présence d'une barrière écologique à la dispersion des gloutons (zone fortement anthropisée). Les flèches indiquent des migrants potentiels entre unités génétiques. (Dessiné d'après [9] et [12]).

5. APPLICATION AUX CHEVREUILS DU SUD-OUEST DE LA FRANCE

Le chevreuil est largement répandu en Europe et est en croissance démographique depuis la seconde moitié du XX^{e} siècle, créant d'importants conflits avec les forestiers, les agriculteurs ainsi que de nombreuses collisions avec les automobilistes [17]. Les populations sont gérées par le biais de plans de chasse et des connaissances sur la biologie de l'espèce sont actuellement nécessaires pour définir au mieux des unités de gestion pertinentes. Des études précédentes ayant révélé une faible différenciation génétique sur des aires géographiques similaires ou supérieures [18], [19] cette application était un bon moyen de tester la performance de notre méthode dans un contexte de très faible différenciation génétique.

Initialement inféodé aux habitats forestiers, le chevreuil a récemment colonisé des espaces agricoles ouverts aux cours des dernières décennies. Dans la plupart des paysages fragmentés, cette espèce reste néanmoins inféodée

aux habitats forestiers [19], [20]. D'autre part, les zones urbaines (villages et routes) semblent réduire les flux de gènes entre les populations [18]. L'aire d'étude est une région agricole principalement d'élevage mais avec des cultures (blé, maïs, sorgho, tournesol) et des habitats forestiers très fragmentés. Elle est traversée par une autoroute grillagée, plusieurs canaux dont les rives très pentues sont bétonnées, et la Garonne (fig. 4), barrières potentielles aux déplacements des chevreuils [16].

Mille cent quarante huit échantillons de tissus, localisés avec une précision de 1 km, ont été collectés entre 2000 et 2004 (fig. 4.A). Nous avons utilisé 11 marqueurs microsatellites spécifiques du Chevreuil selon le protocole décrit par Galan *et al.* [21]. Comme pour le Glouton nous avons d'abord réalisé avec GENELAND un ensemble de simulations dans lesquelles le nombre de populations K est inconnu et donc variable. La valeur de K a été estimée par ces premiers calculs, puis nous avons relancé l'algorithme en fixant K à cette valeur afin d'estimer les autres paramètres (en particulier l'assignation des individus aux populations). Nous avons effectué cinq simulations en permettant à K de varier librement entre 1 et 30 avec les paramètres suivants : 500 000 simulations MCMC, intensité maximale du processus de Poisson fixée à 500, incertitude des coordonnées spatiales fixée à 1km, nombre maximum de noyaux dans la partition de Voronoi fixé à 200. Nous avons utilisé le modèle D pour l'estimation des fréquences alléliques. Nous avons ensuite estimé le nombre de populations par la valeur modale de K dans ces 5 simulations indépendantes, puis lancé 100 simulations MCMC avec K fixé à cette valeur sans toucher aux autres paramètres. Nous avons calculé la moyenne du logarithme de la probabilité à posteriori de chacune de ces 100 simulations et sélectionné les 10 simulations avec les plus fortes valeurs. La probabilité d'appartenance à posteriori de chaque pixel du domaine spatial a été calculée pour chacune de ces dix simulations (après une période de chauffe de 50 000 itérations), de même que l'assignation des individus aux populations inférées. Le nombre de pixels a été fixé à des valeurs élevées, 500 pixels sur l'axe X et 380 sur l'axe Y, de façon à éviter d'avoir deux individus dans le même pixel. Enfin, nous avons examiné la congruence des résultats de ces dix meilleures simulations.

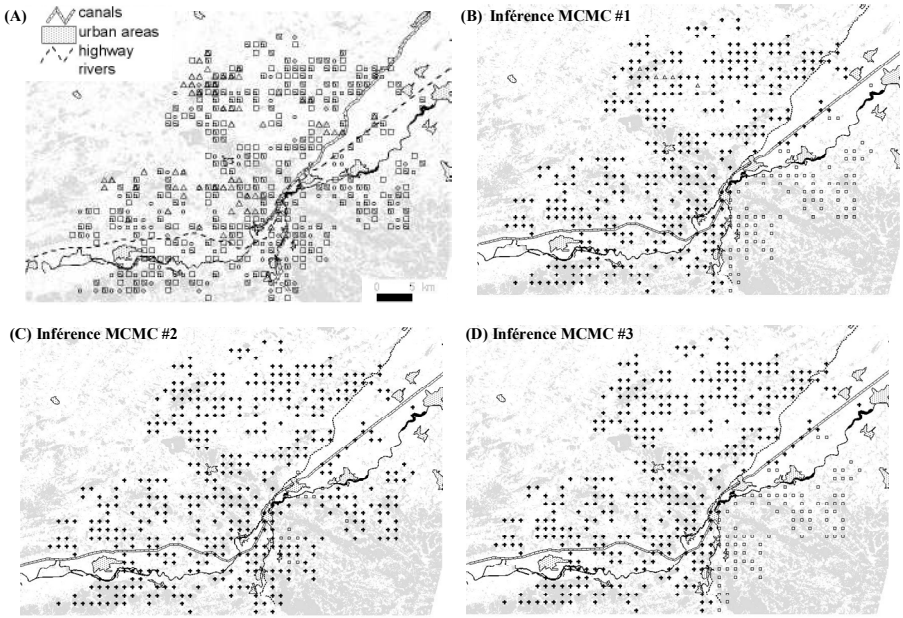

Figure 4 : Structuration génétique du chevreuil (*Capreolus capreolus*) dans le Sud-Ouest de la France. (A) Localisation des échantillons (carrés : mâles ; triangles : femelles) et des principales caractéristiques du paysage (zones boisées en gris) ; (B à D) Assignation génétique avec Geneland par les trois meilleures inférences MCMC (*i.e.* plus fortes valeurs de la moyenne du logarithme de la probabilité à posteriori, voir le texte pour plus de détail). Les sept suivantes sont similaires à l'inférence #3 excepté pour une dizaine d'individus situés entre l'autoroute et la Garonne dans l'Est de la zone qui sont alternativement assignés à une des deux populations. (Dessiné d'après [16]).

Les analyses avec GENELAND donnent différentes valeurs modales du nombre de populations : K = 8, 9, 11 et 14 (deux fois). Étant donné la tendance de GENELAND, évoquée précédemment, à détecter parfois des populations qui ne sont les plus probables pour aucun des individus, nous avons fixé K à 8. Les 10 meilleures simulations parmi les 100 réalisées avec K fixé à cette valeur proposent l'assignation des individus dans deux populations seulement (excepté une simulation, où quelques individus ont été attribués à une troisième population, cf. ci-dessous). Les autres populations inférées sont des populations fantômes, dont les aires correspondent à des zones où aucun individu n'a été échantillonné. Pour huit des simulations, la frontière entre les deux populations est localisée entre le canal et l'autoroute au Nord, et dans une zone comprenant un affluent de la Garonne et des canaux à l'Ouest (fig. 4.D). Deux légères variations à ce patron général sont proposées par les simulations #1 et #2 : la première propose une troisième population constituée de 29 individus au nord de la zone d'étude (fig. 4.B) ;

la seconde propose une partition entre une population principale dans le Nord et une petite population (29 individus) dans le Sud-Est (fig. 4.C). Les statistiques F (Fstat 2.9.3.2 [22]) appliquées aux populations inférées par les huit simulations similaires à la figure 4.D indiquent des valeurs de F_{IS} faibles (0,016 et 0,017 pour les deux populations) et une valeur de F_{ST} faible mais significative (0,008, intervalle de confiance à 95 %, 0,004 -0,012). Une AMOVA (Arlequin [23]) révèle que la majorité de la variation génétique réside au sein des populations (99,24 %), néanmoins la variance entre les populations est très significative ($P < 0,0001$).

Nos résultats suggèrent une rupture de flux de gènes au niveau d'une zone qui comprend une autoroute grillagée, la Garonne, plusieurs canaux et des zones urbanisées [16]. Il est probable qu'aucun de ces éléments ne soit une barrière infranchissable, excepté les canaux, dont les parois bétonnées empêchent les animaux qui y tombent de pouvoir en ressortir. En effet, les chevreuils sont capables de nager sur de longues distances et d'autre part, quelques passages à faune sauvage ont été aménagés sur une portion de l'autoroute. Cependant, ces structures limitent probablement les déplacements des chevreuils et il est possible que leur effet cumulé entraîne la rupture de flux de gènes observée. Cette étude confirme l'impact des réseaux routiers sur la structure des populations naturelles déjà mise en évidence pour le Campagnol roussâtre (*Clethrionomys glareolus* [24]) et le Scarabée terrestre (*Abax parallelepipedus* [25]).

6. CONCLUSION ET PERSPECTIVES

La génétique du paysage est une approche pertinente pour l'étude de la perméabilité des structures paysagères et pour inférer la connectivité d'un paysage. Malgré son potentiel [1] cette approche assez récente est encore peu développée (cf. cependant [19], [26], [27], [28]), notamment en raison d'un manque d'outils statistiques adaptés. Notre étude montre que l'approche géostatistique proposée par GENELAND est capable de détecter et de localiser des ruptures de flux de gènes cohérentes avec les connaissances sur l'écologie des organismes étudiés. D'autre part, nous avons montré que l'approche géostatistique de GENELAND est plus sensible que les approches statistiques non spatiales équivalentes. Dans le cas des gloutons, nous détectons une population supplémentaire à celle de la publication initiale [12]. Dans le cas des chevreuils, les approches non spatiales ne détectent aucune rupture de flux de gènes [16]. Dans les deux cas, les frontières supplémentaires détectées par GENELAND sont validées par les connaissances sur les structures paysagères et l'écologie des déplacements des mammifères étudiés (cf. discussion dans [12], [16]). Ce résultat soutient empiriquement la pertinence de l'information spatiale pour inférer des structu-

res génétiques, notamment dans le cadre d'une très faible différenciation. C'est une information importante car ce type de situation est fréquente dans les problématiques de conservation où les populations sont affectées par des modifications paysagères récentes, et aucune méthode jusqu'à présent n'était capable de traiter efficacement ces cas de figure.

Dans le modèle développé, les individus d'une même population sont « conditionnellement indépendants », c'est-à-dire qu'une fois les paramètres estimés, il n'y a plus de dépendance résiduelle, les paramètres expliquent « totalement » l'écart à l'indépendance observé. Cette hypothèse est restrictive dans un contexte de populations naturelles (nombreuses sources possibles d'écart à l'équilibre de Hardy-Weinberg). L'efficacité et la sensibilité de cette méthode devraient maintenant être testées dans des situations où on observe une dépendance spatiale entre les génotypes d'une même population. Il pourrait être envisagé de mesurer l'influence de l'écart à l'hypothèse de Hardy-Weinberg sur la stabilité et la robustesse de l'analyse statistique en ayant recours à des jeux de données simulées, les seuls qui permettent d'envisager un large éventail de situations et pour lesquelles les attendus sont parfaitement connus. Deux cas de figure couramment rencontrés dans les populations naturelles sont envisageables : « isolement par la distance » et « regroupements spatiaux d'individus apparentés ». Une étape supplémentaire consisterait alors à développer une extension du modèle en injectant une forme de dépendance spatiale entre les individus d'une même population, non prise en compte par l'existence de discontinuités abruptes. La prise en compte de cette dépendance spatiale intra-populationnelle devrait permettre une meilleure sensibilité des analyses statistiques.

REMERCIEMENTS

Ce travail a été financé en partie par le Bureau des Ressources Génétiques.

RÉFÉRENCES

(* signale les publications réalisées dans la cadre du projet financé par le BRG).

[1] Manel S, Schwartz MK, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology and Evolution*, 18 (4), 189-197.

[2] Taylor PD, Fahrig L, Henein K, Merriam G (1993) Connectivity is a vital element of landscape structure. *Oikos*, 68 (3), 571-573.

[3] Moritz C (1994) Defining 'Evolutionarily Significant Units' for conservation. *Trends in Evolution and Ecology,* 9, 373-375.

[4] Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using mutilocus genotype data. *Genetics,* 155, 945-959.

[5] Dawson KJ, Belkhir K (2001) A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetical Reseach*, 78, 59-77.

[6] Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164, 1567-1587.

[7] Corander, J, Waldmann, P, Sillanpää, MJ (2003) Bayesian analysis of genetic differentiation between populations. *Genetics*, 163 367-374.

[8] Corander, J, Waldmann, P, Marttinen, P *et al.* (2004) BAPS2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*, 20 (15), 2363-2369.

[9] Cegelski, C, Waits L, Anderson J (2003) Assessing population structure and gene flow in Montana wolverines (*Gulo gulo*) using assignment-based approaches. *Molecular Ecology*, 12, 2907–2918.

[10] Rueness EK, Jorde PE, Hellborg L *et al.* (2003) Cryptic population structure in a large, mobile mammalian predator: the Scandinavian lynx. *Molecular Ecology*, 12, 2623-2633.

[11] Green J, Hjort ML, Richardson S (2003) Highly Structured Stochastic System. Oxford University Press.

[12] *Guillot G, Estoup A, Mortier F, Cosson JF (2005a) A spatial statistical model for landscape genetics. *Genetics*, 170, 1261-1280.

[13] *Guillot G, Mortier F, Estoup A (2005b) Geneland: A computer package for landscape genetics. *Molecular Ecology Notes* 5 (3), 708-711.

[14] Cornuet JM, Piry S, Luikart G *et al.* (1999) New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics*, 153,1989-2000.

[15] Vázquez-Dominguez E, Paetkau D, Tucker N *et al.* (2001) Resolution of natural groups using iterative assignment tests: an example from two species of Australian native rats (*Rattus*). *Molecular Ecology*, 10, 2069–2078.

[16] *Coulon A, Guillot G, Cosson JF *et al.* (2006) Genetic structure is influenced by landscape features: empirical evidence from a roe deer population. *Molecular Ecology* 15, 1669-1679.

[17] Cederlund G, Bergqvist J, Kjellander P *et al.* (1998) Managing roe deer and their impact on the environment: maximising benefits and minimising costs. In: *The European roe deer: the biology of success* (eds. R. Andersen, P. Duncan & J. D. C. Linnell), Oslo, pp. 189-219.

[18] Wang M, Schreiber A (2001) The impact of habitat fragmentation and social structure on the population genetics of roe deer (*Capreolus capreolus L.*) in Central Europe. *Heredity*, 86, 703-715.

[19] Coulon A, Cosson JF, Angibault JM *et al.* (2004) Landscape connectivity influences gene flow in a roe deer population inhabiting a fragmented landscape: an individual-based approach. *Molecular Ecology*, 13, 2841-2850.

[20] Hewison AJM, Vincent JP, Joachim J *et al.* (2001) The effects of woodland fragmentation and human activity on roe deer distribution in agricultural landscapes. *Canadian Journal of Zoology*, 79, 679-689.

[21] Galan M, Cosson JF, Aulagnier S *et al.* (2003) Cross-amplification tests of ungulate primers in roe deer (*Capreolus capreolus*) to develop a multiplex panel of 12 microsatellite loci. *Molecular Ecology Notes*, 3, 142-146.

[22] Goudet J (2001) FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3). Available from http://www.unil.ch/izea/softwares/fstat.html.

[23] Schneider S., Roessli D., Excoffier L. (2000) Arlequin ver. 2.000: A software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland.

[24] Gerlach G, Musolf K (2000) Fragmentation of landscape as a cause for genetic subdivision in bank voles. *Conservation Biology*, 14 (4), 1066-1074.

[25] Keller I, Nentwig W, Largiadèr CR (2004) Recent habitat fragmentation due to roads can lead to significant genetic differenciation in an abundant flightless ground beetle. *Molecular Ecology*, 13, 2983-2994.

[26] Michels E, Cottenie K, Neys L *et al.* (2001) Geographical and genetic distances among zooplankton populations in a set of interconnected ponds: a plea for using GIS modelling of the effective geographical distance. *Molecular Ecology*, 10, 1929–1938.

[27] Funk CW, Blouin MS, Corn PS *et al.* (2005) Population structure of Columbia spotted frogs (*Rana luteiventris*) is strongly affected by the landscape. *Molecular Ecology*, 14, 483-496.

[28] Spear SF, Peterson MD, Matocq MD, Storfer A (2005) Landscape genetics of the blotched tiger salamander (*Ambystoma tigrinum melanostictum*). *Molecular Ecology*, 14, 2553–2564.